Is a Website or E-Learning Product Working Well? How Many Users Should You Test?

Theodore Frick,[1] Tyler Dodge,[2] Xiaojing Liu,[3] Bude Su[4]
Indiana University, Bloomington

[1] Correspondence about this article should be addressed to
Theodore Frick, Department of Instructional Systems Technology, School of Education,
Room 2218, 201 N. Rose Ave., Bloomington, IN 47405. 812/856-8460, or by e-mail, frick@indiana.edu

[2] Department of Instructional Systems Technology, Indiana University
812/856-8770, tdodge@indiana.edu

[3] Department of Instructional Systems Technology, Indiana University
812/334-7622, xliu@indiana.edu

[4] Department of Instructional Systems Technology, Indiana University
812/857-0407, subude@indiana.edu

Abstract

How can one determine *efficiently* if an informational website or an e-learning product is working well? Relatively small numbers of the target audience are needed to improve a product during formative evaluation and usability testing as part of product development and revision cycles. However, during summative evaluation, how many subjects are needed to determine product effectiveness?

When investigating the number of subjects needed for usability tests, a Poisson probability model was found to be a reasonable fit to extant data (Nielsen & Landauer, 1993; Virzi, 1990, 1992). However, this model was chosen on the basis of the number of subjects needed to identify important usability problems with a product, *not* for determining its effectiveness. To determine if a Website or e-learning product is working well, we investigated the predictive validity of a discrete Bayesian decision model: the Sequential Probability Ratio Test (SPRT) -- originally developed by Wald (1947). Fifty-one people representing a campus community participated in a usability test of the university library online catalog search tool, and the results were analyzed *post hoc* with SPRT re-enactments to simulate sequential decision making after testing each subject. Across a range of parameters, the Bayesian SPRT reached the same conclusion as reflected by the entire sample with many fewer subjects, utilizing typically small Type I and II error rates. The study provides evidence of the usefulness of the SPRT decision model in situations where determination of effectiveness is the goal (product works well or not). The SPRT maximizes efficiency by testing only as many users as necessary to reach a confident conclusion.

Introduction

When investigating the number of subjects needed for usability tests, a Poisson probability model has been found to be a reasonable fit to extant data (Nielsen & Landauer, 1993; Virzi, 1990, 1992). In perhaps the most contemporary review of issues relating to usability testing, Turner, Nielsen, and Lewis (2002) identified two central concerns: the reliability of traditional testing procedures, and the validity of the traditional model of problem detection. More specifically, regarding the formula used for estimating problem detection, they questioned whether the probability of a problem being detected can be modeled fairly with a unitary probability value.

Moreover, this model was chosen on the basis of the number of subjects needed to identify important usability problems with a product, *not* for determining its effectiveness. The purpose of the current study is to offer an approach to usability testing utilizing the Sequential Probability Ratio Test (SPRT) to determine product effectiveness (Wald, 1947). Rather than testing with a predetermined sample size, SPRT analyzes the knowledge accumulating during testing to determine when to stop testing, significantly reducing the number of subjects required. Wald's sequential probability ratio test (SPRT) went beyond the work of Thomas Bayes, who was concerned about how decisions can be reached as evidence accumulates. Wald's SPRT gives us rules for when to stop collecting evidence and reach a conclusion. The SPRT also tells us the likelihood that we would be reaching a wrong conclusion. The SPRT was originally used for manufacturing quality control decisions, and was considered so important that it was classified as a defense secret by the U.S. government during World War II.

Usability testing traditionally serves one of two purposes, either formative or summative evaluation, and the contrasting goals of these two forms of evaluations are reflected in approaches to usability testing as either problem detection or determining effectiveness. Most of the literature concerns problem detection, and a central tenet is that, given enough users and evaluators, most if not all of a product’s usability problems may be uncovered. Of course, when ungainly numbers would be needed, a balance must struck between investment in usability testing and returns on investment, that is, identified problems. Problem detection studies traditionally use the probabilistic Poisson model to determine the number of subjects needed.

$$\textit{Uncovered Problems} = N(1 - (1 - ?)^{n})$$

N: total number of usability problems in the design
?: proportion of usability problems discovered while testing a single user
n: number of subjects

Given an accurate probability estimate, this simple formula provides a fairly good prediction of the number of subjects needed to determine certain proportion of usability problems. Offering the first evidence supporting use of the model, Virzi (1992) found that observing four or five users would reveal 80% of a product’s usability problems, but this estimate and a host of related issues have been actively debated over the last decade. The accumulation of literature relating to problem detection has raised doubts regarding the certainty of the “five users” rule, as well as bringing to light several previously unrecognized issues relating to usability testing, including the probability of error detection, the assumption of homogeneity among users, the inconsistency between evaluators, and the definition of the usability task.

The first central issue relates to the probability of detecting a problem during testing. An average value of between .30 and .40 was suggested by a number of studies (Nielsen & Landauer, 1993; Virzi, 1990, 1992) and, based on the cumulative binomial probability formula, led to the statement that testing only four or five users will uncover 80% of the usability problems. Indeed, the diminishing returns after testing five users, a rule-of-thumb popularized in Nielsen's (2000) online *Alertbox*, continues to gain acceptance. While the rule holds true for probabilities in that range, other studies suggest that the actual probability of finding usability problems may be considerably lower (Lewis, 1994), with the result that usability testing would require a significantly greater number of users. For the *p* value of .16 that Lewis found, fully twice as many users would be needed to find 80% of the problems. Further, though Virzi asserted that the more severe problems would generally be identified before those of lesser import, Lewis found no such correlation; indeed, findings by Spool and Schroeder (2001) likewise challenge Virzi's claim, indicating that testing with a small number of users could be problematic for products with potentially hazardous problems.

Not only challenging the accepted sample size, concern over the probability levels of error detection has brought other issues to the discussion of usability testing. To begin, Caulton (2001) concluded that the assumption of homogeneity among users—the equal likelihood of all users to encounter all problems—not only accounts for the discrepancy between Lewis and his predecessors but compromises usability findings based on the assumption. Virzi's (1992) binomial model, Caulton explains, assumes homogeneity among the subjects, who "must be equally likely to encounter *all* problems" (p. 2). By introducing two classes of usability problems (common and rare) into the model, Caulton duplicates Lewis' (1994) findings that rare problems are not likely to be detected with only five subjects. Moreover, Caulton shows that heterogeneous subgroups likewise create the need for increased numbers of users to detect the same number of usability problems. Further, Caulton's conclusion accounts for the assumption by Virzi (1992), uncorroborated by Lewis (1994), that the probability of detecting a problem is positively correlated to the severity of the problem: "it is possible that *p* and severity *were* correlated in Lewis' data, but that subgroups masked the correlation" (p. 6). In this way, the discrepancy between Virzi and Lewis may be explained, but only by introducing the complex issue of user group composition into usability testing.

The problems associated with the homogeneity assumption were also put forth by Woolrych and Cockton (2001), who, like Caulton (2001), challenged the validity of Nielsen and Landauer's (1993) formula supporting their claim that five users are enough to detect the majority of usability problems. First, through a discussion of statistical theory, the authors showed that the probability of errors being found may be much lower than is fixed in the formula. To demonstrate their claim, they cite Spool and Schroeder's (2001) study in which goal-oriented testing drove the probability much lower than Nielsen and Landauer's 31%. Then, citing their own study of heuristic evaluation, the authors show that the probability of error detection depends not only on the severity of the problem but on differences between users, the same issue explicated by Caulton.

Just as different users encounter different usability problems, so do different evaluators identify the problems inconsistently, a pattern referred to as the evaluator effect (Hertzum & Jacobsen, 2001; Jacobsen, Hertzum, & John, 1998). In these studies and others (Molich et al., 1998), it was found that even when employing similar evaluation methodologies to test the usability of identical products, evaluators differ in their assessment of which observations constitute usability problems. The subjective and inconsistent identification of problems, even when using such relatively strict usability evaluation methods as cognitive walkthroughs and think aloud procedures among experienced professionals, lead to inter-evaluator agreement as low as 5% to 65%. On the one hand, this suggests that testing with multiple evaluators will uncover more and more varied problems than with a single evaluator, and indeed, Jacobsen, Hertzum, & John (1998) note that "the effect of adding more evaluators to a usability test resembles the effect of adding more users" (p. 256). On the other hand, the disparity among evaluators problematizes the "apparent reality of usability improvement achieved through iterative application of usability evaluation methods" (Lewis, 2001, p. 346).

In an article cited above, Spool and Schroeder (2001) reveal a fourth issue central to the question of the number of users, namely the definition of the usability task. In contrast to Nielsen and Landauer's testing with clearly defined tasks, or what Hudson (2001) calls "task-directed" testing, Spool and Schroeder allowed users to define their own goals, or "goal-directed" testing. That is, the five-user rule relates to situations in which all users engage in the same tasks of the product under evaluation, but when testing entails authentic users engaged in authentic tasks, the probabilities of error detection may be no higher than .16; at such low levels, the number of users Spool and Schroeder found necessary may range from around six to over thirty. Task-directed testing cannot achieve the coverage that goal-directed testing does, and authentic website use certainly entails a number of personal decisions, but the author's methodology directed the users to conduct a purchase which arguably does not fairly characterize the majority of tasks, even on commercial sites.

While a major goal of usability test is to identify design problems and recommend changes for a certain product, it is equally important to verify whether a product is working well enough that there is no need to invest additional resources on re-design and further evaluation; moreover, in the case of product effectiveness, we concern ourselves only with the effective use of the product, not the insights coming from testing for further development. While a considerable number of research studies address the identification of usability problems, our research team was unable to identify any significant literature addressing the number of users needed to conclude if a product is working well enough to stop further testing. Perhaps the simplest and most intuitive method is a simple calculation of success rate, or the percentage of successes encountered during usability testing. As Nielsen (2001) explains, success rates "provide a general picture of how [a product] supports users" and represent "the bottom line of usability" (n.p.), but beyond an explanation of the usefulness of tallying partial successes, he does not discuss such implications as the statistical limitations of such a metric.

The purpose of the current study is to employ the Sequential Probability Ratio Test (SPRT) (Wald, 1947) to determine the number of subjects needed to conclude whether or not a website meets a given effectiveness criterion threshold. Under the framework of classical hypothesis testing, the number of subjects needed to test a product's effectiveness can be predicted with the specification of acceptable levels of Type I and Type II errors, along with the population variance. Significantly, a relatively large sample size is usually needed to conclude whether the findings are generalizable. In this study, we propose an alternative approach using Bayesian reasoning to determine number of subjects needed in usability testing. Wald's (1945) SPRT offers an elegant framework for making statistical decisions between different courses of action. Through SPRT, one can use prior probabilities to express the preference for one or the other action and determine how these beliefs can change based on the accumulation of knowledge from observed data (Wald, 1945). Wald (1945) claimed that using SPRT to make a sampling plan leads to an average saving of at least 48% in the necessary number of observations, compared with the classical hypothesis testing. Later Colton and McPherson (1976) similarly found that using SPRT can achieve potential economy by testing fewer samples than fixed-sample-size while still attaining the desired level of statistical significance.

SPRT is a methodology for deciding between two alternatives under sequential observations. Though not developed under the framework of Bayesian reasoning, SPRT can be regarded as an extension of Bayesian theorem with addition of stopping rules (Frick, 1989). The central tenet of Bayes' theorem is its likelihood principle: *posterior probability* is proportional to *the prior probability* multiplied by the *likelihood* of that alternative, which can be expressed as follows:

Posterior probability ∝ Prior probability × Likelihood

Likelihood is the conditional probability of the event when a particular alternative is true. Prior probability is people's prior knowledge about the probability distribution of the alternative before the observation; after the observation, people's beliefs in the alternative will change due to likelihood principle. Posterior probability is people's knowledge about the probability distribution of the alternative after observation (Schmitt, 1969). When the observations are made sequentially, the posterior probability of one observation becomes the prior probability of the next observation. When several alternatives involved, the Bayesian theorem can be expressed as follows:

If

i. Alternatives are mutually exclusive and exhaustive;
ii. Let $P_0 (A_i)$ be the prior probability of A_i;
iii. X is the observation;
iv. $P (X | A_i)$ is the probability of the observation given that A_i is true.

Then the probability of A_i is

$$P (A_i | X) = \frac{P_0 (A_i) P (X | A_i)}{? P_0 (A_i) P (X | A_i)} \qquad (1)$$

Assuming we have to decide between two alternatives with error probabilities of a (Type I) and ß (Type II), by using Bayesian reasoning to compute the posterior probability of the alternatives after each observation, at a certain point the results will be highly in favor of one alternative over the other. SPRT offers stopping rules that can be used to cease observation and reach a conclusion with a and ß error tolerances. The rules are as follows:

Rule 1: Compute the ratio (PR) of the posterior probabilities of the alternatives. If PR is greater than or equal to (1 – ß) / a, then choose the first alternative;

Rule 2: If PR is less than or equal to ß / (1 – a), then choose the second alternative;

Rule 3: If neither Rule 1 nor Rule 2 is true, then another observation is needed. After a new result obtained, then update the posterior probabilities and reapply the three rules.

Let us apply this rule in the context of the present study: we want to decide between the hypotheses that either the website is effective (I) or the website is not effective (II). In this case, we need to make a sequence of observations in the context of usability testing to determine which option to choose. After randomly selecting a subject from the population using the web site, we calculate the probability ratio, PR:

$$PR = \frac{P_{e0}\, P_e^{\,s} (1 - P_e)^f}{P_{n0}\, P_n^{\,s} (1 - P_n)^{\,f}} \qquad (2)$$

In Equation 2, P_{e0} and P_{n0} are the initial probabilities of effectiveness and non-effectiveness, respectively. P_e represents the probability of randomly selecting a subject that would complete the usability tasks if the website is effective, and P_n represents the probability of randomly selecting a subject that would fail in the tasks even if the website is effective; a is the error probability of concluding the website is effective when it is actually not effective, and ß is the error probability of concluding that the website is not effective when it actually is.

In Equation 2, we can assume P_{e0} and P_{n0} to be equal so that they can cancel each other out in the equation. Assuming that *s* and *f* refer to the numbers of users who are successful or not at completing tasks, we have contextualized stopping rules as follows:

Rule 1: If PR = (1 – ß) / a, then stop the testing and conclude that the website is effective;

Rule 2: If PR = ß / 1 – a, then stop the testing and conclude that the website is not effective;

Rule 3: If (1– ß) / a < PR < (1– ß) / a, then randomly select another subject and test again, increment *s* or *f* accordingly, recalculate PR, and apply Rule 1 to Rule 3 again.

This kind of reasoning assumes a number of necessary prerequisites, some of which are likewise assumed in inferential statistics:

1. Observations must be independent. That is, the outcome of one observation should not influence the outcome of another.
2. Observations must be randomly sampled. Random sampling is necessary for generalizing results from sample to population.

In addition, we make the following assumptions:

3. Alternatives must be mutually exclusive and exhaustive. (Though the stopping rules of SPRT may be applied to more than two alternatives, in this study, we consider only two alternatives, namely effective or not effective.)
4. The conditional probability of each alternative must be specified.

Because of its potential, SPRT has been widely applied in industry to test the quality of manufactured products, and in education too, Bayesian procedures have been used in computerized adaptive testing (CAT) to make mastery and non-mastery decisions. Studies have shown that the SPRT can be successfully applied in CAT using item response theory (IRT) (Frick, 1989; Frick, 1992; Lewis & Sheenan, 1990; Reckase, 1994). Frick (1989) argued that though SPRT does not take into account variability in item difficulty, discrimination, and guessing factors, the decisions of mastery or non-mastery reached by SPRT in his study agreed very highly with those reached through administering the entire item pools to examinees. Frick concluded that because of its simplicity and practicality, SPRT offers a viable model to achieve reliable results in CAT, provided that the method is used conservatively (e.g., small error probabilities). Like determining mastery or non-mastery of a educational content area, the task of determining site effectiveness is a binary decision; moreover, the task of determining site effectiveness with the fewest subjects possible is similar to the task of determining mastery by sampling as few test items as possible, thus warranting our application of SPRT to usability testing. In this research, we seek to establish whether the SPRT has predicative validity in reaching reliable conclusions as to a website's effectiveness using as few subjects as possible.

Method

A total of 51 people 18 years or older participated in this study at a large mid-western university and its community. The subjects were recruited through a method of stratified convenience sampling. First, we identified five strata of users of the University library resources: undergraduate students, graduate students, faculty, staff, and non-university affiliated community members. In order to have a sufficient number of subjects for SPRT analysis, we determined that we needed about 50 subjects, and to obtain a sample corresponding to the population demography of the university, we determined to seek the following proportions: 33 undergraduate students, 11 graduate students, 3 faculty, 2 staff, and 2 community members. Subjects in all strata were obtained according to convenience (discussed below) in precisely these proportions, with the exception that one additional undergraduate subject was tested.

Among the research participants, 5 subjects reported that they used the university library's online catalog often, 18 subjects used it occasionally, 20 seldom used it , and 8 had never used it at all. In addition to self-reported usage of the online catalog, subjects were asked to report their confidence using other similar searches. Specifically, asked to respond to the statement "I am confident using search engines" in terms of a 5-point Likert scale, 15 subjects strongly agreed, 27 subjects agreed, 7 subjects were neutral or undecided, 1 subject disagreed, and 1 subject strongly disagreed with the statement of confidence.

This research involved testing the usability of the online catalog of the Indiana University libraries (IUCAT). Determining the success of IUCAT, while of interest to stakeholders in the site's usability, remained of secondary interest after our primary question regarding the applicability of SPRT in determining the number of users necessary to determine success. Accordingly, rather than engage in the more thorough but challenging style of testing involving goal-oriented tasks, we focused our testing on particular tasks addressed by the catalog.

In order to provide some empirical basis for our task selection, we consulted the documentation relating to the usability testing of another university's online catalog, namely the study conducted by the Institute of Museum and Library Services of the University of Texas at Austin (2001). In one phase of their study, the researchers conducted focus groups with volunteers recruited from the University libraries staff; nearly three-fourths of the volunteers, being librarians from the public services cluster, were asked to represent "those library users who are served by the Web site and with whom the professional staff has contact on a regular basis" (Institute of Museum and Library Services [IMLS], 2000b). These librarians were asked to "think of a task that you typically do on UT Library on Line" and to "briefly describe this task" (IMLS, 2000a), and we coded the list tasks to identify the most prevalent among them: finding details on a specific book, and finding materials on a specific topic, including searches of works by a given author.

From these categories we developed our tasks, which involved (1) identifying the most recent book in the library system written by a specific author, and (2) determining to which library or libraries a specific book belongs. These two tasks involve many of the same procedures as other tasks we did not test; they entail use of many of the same features of the site, and they require many the same skills on the part of the user. We believe, therefore, that these two tasks are representative of most if not all of the other tasks addressed by the online catalog, and so we operationally define the catalog's success in terms of typical users' successful completion of these two tasks.

Testing proceeded in the following manner. After identifying the campus buildings with the greatest number of computer laboratories available for student use, we visited the laboratories in their rank order, on different week days, and at various times of day. When the laboratories were crowded, we solicited students waiting in line; otherwise, we solicited them at their workstation, working systematically through the laboratory. No more than eight subjects were recruited from any single laboratory, and no more than ten on any single day. Faculty and staff were solicited in a similar manner: we identified the schools with the most students and visited the buildings on different days and at different times; we positioned ourselves at a haphazard location in the building and systematically solicited faculty and staff at their desks. The community member was chosen by convenience and tested at home. Two researchers from the team conducted each usability test, one facilitating the testing procedures, and the other recording observations regarding the subject's activities during the completion of the two designated tasks. In addition, the subjects completed a brief questionnaire of their computer experience and background information. Testing proceeded in this way until the target samples were satisfied. The majority of the computer workstations featured Windows operating systems, though a small number of Macintosh machines were also used in the testing; all of the testing employed the Microsoft *Internet Explorer* software browser.

The SPRT analysis proceeded as follows. First, using *Statistical Package for the Social Sciences* (SPSS) version 11.0, we analyzed the descriptive statistics relating to subject background information, time spent on each task, and successfulness of the usability tasks. Second, we used a random number table to randomize the record order of the usability test data. As mentioned above, we had collected data using from four to eight subjects from each computer cluster and had labeled the records chronologically; the purpose of randomizing the record order was to avoid possible bias relating to the data collection procedure. Third, the data records were individually coded as either success or failure based on how well the subject had performed the tasks: specifically, if a subject succeeded on both tasks, this counted as a success case, but if a subject failed on both tasks, failed on either one of the two tasks, or only partially succeeded on one or both of the tasks, we coded it as a failure case. Forth, we used the SPRT simulation coded by Frick (2001) to analyze how many subjects would be needed to conclude whether the online catalog is effective or not. Finally, changing various parameters of the SPRT, we compared the number of subjects needed to determine effectiveness reached by different criteria.

Results

We first defined the SPRT parameters as follows. If the online catalog website is effective, we would expect 90% or more of the users to succeed in the tasks set to them. If the success rate is 60% or less, we would conclude that the site is not effective. In other words, we are presented with two alternatives: (1) the website is effective, and (2) the web site is not effective. Stated as conditional probabilities, we have this:

Probability (success | website is effective) = .90 or higher {1}
Probability (success | website ineffective) = .60 or less {2}
(a=.05, ß=.05)

The first randomly selected subject from the pool of 51 subjects did not pass the test (this person failed on the second task). Thus, to this point we have observed one failure and no successes. The results are summarized in Table 1.1. The posterior probability was inclined toward determining the site's effectiveness as a failure at a .80 probability level, while the posterior probability of success was only .20. Still, SPRT could not make a conclusion at this time.

Table 1.1
SPRT Results after 1 Subject

	Probability			
Alternative	Prior	Conditional	Joint	Posterior
Success	.5000 ×	.1000 =	.0500 / sum =	.2000
Failure	.5000 ×	.4000 =	.2000 / sum =	.8000
			sum = 0.2500	

The second randomly selected subject succeeded on both tasks, soaltogether we have observed one success and one failure. The SPRT results are presented in Table 1.2.

Table 1.2
SPRT Results after 2 Subjects

	Probability			
Alternative	Prior	Conditional	Joint	Posterior
Success	.2000 ×	.9000 =	.1800 / sum =	.2727
Failure	.8000 ×	.6000 =	.4800 / sum =	.7273
			sum =.6600	

After this turn, the posterior probability for failure dropped from .80 to approximately .73, and the posterior probability for success increased from 0.20 to approximately 0.27. Similar steps were repeated until the 12th subject was tested. Through this round of analysis, still only one subject—the first—had failed the tasks; the remainder were successes. The SPRT results at this point appear Table 1.3.

Table 1.3
SPRT Results after 12 Subjects

	Probability			
Alternative	Prior	Conditional	Joint	Posterior
Success	.9351 ×	.9000 =	.8416 / sum =	.9558
Failure	.0648 ×	.6000 =	.0389 / sum =	.0442
			sum = 0.8805	

With this turn, the posterior probability for success had risen sufficiently to make a determination:

.956 / .0442 = 21.629 = (1 – ß) / a = 19

Accordingly, we aborted testing and concluded that the website is effective.

In summary, the SPRT analysis of subjects from our sample pool in a random order allowed for success to be determined with 12 subjects. This result, of course, reflects the input data, which included one failure case among the successes; significantly, the failure could have appeared in any position among the first eight iterations of the SRPT with the same result. Given the same parameters and without a failure case among the first 8 entries, SPRT would have reached the conclusion of overall success with only 8 subjects, and conversely, if all of the initial entries were failure cases, SPRT would have determined overall failure with only 3 subjects.

In order to test the predictive validity of SPRT, we analyzed the same random data set under different conditions. Table 2.1 lists the results of running SPRT while keeping Alpha and Beta error constant (a=.05, ß=.05) but changing the success and failure rates.

Table 2.1

SPRT Test Results with a=.05, ß=.05

Rate		Observation			
Success	Failure	Success	Failure	Total Users	Conclusion
90%	50%	8	1	9	Success
90%	60%	11	1	12	Success
90%	70%	17	1	18	Success
90%	80%	46	5	51	No conclusion

The results reveal that as the success and failure settings became closer together, more subjects were needed to reach the conclusion regarding website effectiveness (for example, if the failure rate increased from 60% to 70%, SPRT required an additional 6 successful users to conclude the overall effectiveness of the website); moreover, as the difference between the success rate and failure rate approaches zero, the number of users required to reach a conclusion becomes exponentially large.

On the other hand, we could keep the success and failure rates constant and reduce the Alpha and Beta error levels. Table 2.2 displays the results of running SPRT while maintaining Alpha and Beta error levels (a=.05, ß=.05).

Table 2.2

SPRT Test Results with Success Level=90%, Failure Level=60%

Level		Observation			
a	ß	Success	Failure	Total Users	Conclusion
0.05	0.05	11	1	12	Success
0.03	0.03	12	1	13	Success
0.01	0.01	15	1	16	Success

These results reveal that as the Alpha and Beta error levels were reduced, more subjects were needed to reach a conclusion regarding website effectiveness (e.g., to reduce the Alpha and Beta error levels from .5 to .1, SPRT required an additional 4 successful users to reach a conclusion). These results demonstrate that SPRT can provide a relatively predictable method of estimating the number of users needed to determine website effectiveness.

Success and failure levels like those above may be common in educational settings, such as computer adaptive testing, but in commercial and industrial contexts—especially ones with high stakes, as in medical and military production—both success and failure levels are likely to be much higher: effective manufacture of pharmaceutical products, for example, may be as high as 99%, while manufacture may be deemed unsatisfactory even at levels as high as 95%. SPRT can be applied in these contexts as well, though, as seen above, a significantly larger sample size will be required to reach a conclusion. As displayed in Table 2.3, running our sample (n=51) through SPRT with levels like mentioned above results ina determination of non-success. As the success and failure settings get closer, more samples were needed to reach a conclusion.

Table 2.3

SPRT Test Results with a=.05, ß=.05

Rate		Observation			
Success	Failure	Success	Failure	Total Users	Conclusion
98%	90%	31	4	35	Non-success
99%	90%	24	3	27	Non-success
99%	98%	41	5	46	Non-success

Discussion

In this study of website effectiveness, the usability data were analyzed by SPRT to determine whether the site was successful or not. Across a range of parameters including error tolerance and thresholds of success and failure, SPRT reached the same conclusion as reflected by the entire sample set, but SPRT required fewer samples to do so. For example, when we continued applying SPRT to our entire set of 51 samples, we reached the same conclusion of success as reached with only 12 subjects. This shows that SPRT bears predictability and reliability in its determinations.

The study provides good evidence of the usefulness of SPRT in usability testing: in summative evaluations or situations where determination of effectiveness rather than error detection is the goal, SPRT provides a method of data analysis with considerable flexibility. Even when performing a single rather than sequential or iterative calculation of success, SPRT affords a simple and sound alternative to raw percentages or statistical procedures such as beta distributions, which are likely to require more users for the same error rates. Moreover, when used sequentially to analyze usability data, SPRT can provide determinations at a substantial reduction of number of users.

At first glance, the requirement for analysis parameters such as success and failure rates may seem to be a limitation of SPRT, but in fact, the beta distribution likewise calculates results according to a cutoff rate, which is, in effect, the average of the success and failure rates; indeed, it may be considered advantageous that SPRT allows success and failure to be specified independently. Accordingly, these parameters should not be considered a fallibility left to the discretion of the analyst but, rather, an opportunity for the testing to reflect the needs of the stakeholders.

While not bearing upon the usefulness of SPRT in usability testing, a few points regarding the actual testing deserve to be mentioned. First, the percentage of failures encountered during the study needs qualification. In several cases, despite the subject's entry of the correct information using the correct submission procedures (e.g., conducting a "title search" of "all libraries"), the server produced incorrect results, that is, results inconsistent with the results produced under the same conditions at other times; despite the fact that the subject used the online catalog in the correct manner, we tallied this as evidence against the site's effectiveness. Further, in most of the testing situations, the subjects experienced inordinate server delays in receiving results; many subjects interpreted this as an error on their part and returned to the search page to review their input, or repeatedly clicked the submit buttons, or in other ways disrupted the original usage scenario. In every case, we let the encounter proceed to its conclusion—often to success, however slow, but in several cases converting what would have been a successful case to one of failure to accomplish the task. Not only did the server, through its errors and delays, contribute to the number of unsuccessful searches, but our own criteria for success may be regarded as unduly stringent. Specifically, onlyif a subject succeeded on both of the tasks did we regard the case as a success; if the subject was successful on one task but only partially successful on another, we counted the entire case as a failure—a definition of success perhaps not reflective of the website owner's own, but one that ultimately provided data suitable to SPRT analysis.

A second consideration is the inconsistency of the appearance of the search page indifferent situations. Specifically, the html coding of the search page specifies that, in the drop-down list from which the user selects which libraries to include in the search, the default or selected option is "all campus libraries," meaning all libraries on the local campus but excluding all libraries on other campuses. In common settings, this default setting is used to guide the search, unless the user selects otherwise, but in the computer laboratories available for student use, this default is overridden: the browser instead presents "all libraries," that is all libraries on all campuses, as the default. This variation resulted in inconsistencies among results. Since the participants were solicited by convenience, their investment in the testing was likely only casual, and indeed, while the tasks were commonplace, they were not intrinsic. As a consequence, though both of the tasks called for the subject to find a reference from any of the libraries within the university system, one of the tasks addressed an item located only in an off-campus library. On this task, then, users at computers other than the campus laboratory workstations would have had to change the option relating to library selection to retrieve the same results as users in the laboratories, that is, to find the correct reference; otherwise, a different result would consistently be returned by the search engine. We considered this difference to be a limitation of the testing procedures (e.g., subjects recruited without compensation) rather than a limitation of the website (though the default option bears significant implications on the usability of the system), and accordingly, for users whose default setting covered only campus libraries, we accepted the alternate answer as correct. As with the errors discussed above, this limitation may bear upon accepting the findings as representative of the site's usability, but not upon the usefulness of SPRT procedures in usability testing more generally.

Finally, a related consideration is the limitation of generalizing from the usability tasks to the catalog search engine more broadly. While several of the features were not tested directly (e.g., searches for journal titles), we nonetheless consider them to be similar in presentation and functionally to the tasks covered by the testing. Accordingly, we may tentatively generalize the site's effectiveness on the tasks tested to reflect the site's effectiveness for the related tasks. Still, this step is problematized by the interaction between the tasks and the libraries searched, but again, this does not pertain to the SPRT analysis.

While the study offers data regarding the usability of a particular website search engine, and while the methods and usability results may inform future studies of website effectiveness, the chief contribution of this study is the demonstration of SPRT's application in usability testing. Further studies may likewise contribute to this body of knowledge through several avenues of inquiry: they may continue comparing SPRT to other statistical procedures to establish its benefits and limitations; explore the range of applications of SPRT to gauge its usefulness and flexibility; and establish methods of implementing SPRT during testing to determine when to stop testing. It is hoped that the present study demonstrates the promise of such pursuits.

References

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1-7.

Colton, T., & McPherson, K. (1976). Two-stage plans compared with fixed-sampling-size and Walt SPRT plans. *Journal of the American Statistical Association, 71.*

Frick, T. W. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, *5*(1), 89-114.

Frick, T.W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

Frick, T. W. (2001). Making Decisions Using Bayesian Reasoning. Retrieved November 30, 2002, from http://education.indiana.edu/~frick/decide/start.html

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 421-443.

Hudson, W. (2001). How many users does it take to change a web site? *SIGCHI Bulletin*, May/June 2001. Retrieved October 8, 2002, from http://www.syntagm.co.uk/design/articles/howmany.htm

Institute of Museum and Library Services (2000a). Evaluation of Effective Library Sites. *Characterization of the use and usability of a web-based digital library* (Usability Testing section). Retrieved October 6, 2002, from http://imls.lib.utexas.edu/usability/evalmit.html

Institute of Museum and Library Services (2000b). Report Covering Project Activities to 6/30/2000. *Characterization of the use and usability of a web-based digital library* (Reports section). Retrieved October 6, 2002, from http://imls.lib.utexas.edu/report/activities00-06-30.html

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. In C.-M. Karat & A. Lund (Eds.), *Human Factors in Computing Systems CHI'98 Summary* (pp. 255-256). New York: ACM Press.

Lewis, C., & Sheenan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 376-386.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, *36*(2), 368-378.

Lewis, J. R. (2001). Introduction: Current issues in usability testing. *International Journal of Human-Computer Interaction*, *13*(4), 343-349.

Littlewood, B., & Wright, D. (1997). Some conservative stopping rules for the operational testing of safety-critical software. *IEEE Transactions on Software Engineering*, *23*(11).

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D. et al. (1998). Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association 1998 (UPA98) Conference* (pp. 189-200). Washington D.C.: Usability Professionals Association. Retrieved October 8, 2002, from http://www.dialogdesign.dk/tekster/cue1/cue1paper.doc

Nielsen, J. (2000). Why you only need to test with 5 users. *Alertbox*, March 19, 2000. Retrieved October 8, 2002, from http://www.useit.com/alertbox/20000319.html

Nielsen, J. (2001). Success rate: The simplest usability metric. *Alertbox*, February 18, 2001. Retrieved November 13, 2002, from http://www.useit.com/alertbox/20010218.html

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of INTERCHI '93* (pp. 206-213). Amsterdam, The Netherlands: ACM Press.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–255). New York: Academic Press.

Schmitt, S. A. (1969). *Measuring uncertainty: An elementary introduction to Bayesian statistics*. Reading, MA: Addison Welsley Publishing Company, Inc.

Spool, J., & Schroeder, W. (2001). Testing web sites: Five users is no where near enough. In J. Jacko & A. Sears (Eds), *Conference on Human Factors in Computing Systems: CHI 2001 Extended Abstracts* (pp. 285-286). Seattle, WA: ACM Press.

Turner, C. W., Nielsen, J., & Lewis , J. R. (2002). Current issues in the determination of usability test sample size: How many users is enough? In*Usability Professionals' Association 2002 Conference Proceedings*(n.p.). Chicago: UPA.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Human Factors and Ergonomics Society 34th Annual Meeting* (pp. 291-294). Santa Monica, CA: Human Factors and Ergonomics Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*(4), 457-468.

Wald, A. (1945). Sequential method for deciding between two courses of action. *Journal of the American Statistical Association, 40*(231), 277-306.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley & Sons, Inc.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), *Proceedings of IHM-HCI 2001 Conference: Vol. 2* (pp. 105-108). Toulouse, France: Cépadèus Éditions. Retrieved October 8, 2002, from http://www.cet.sunderland.ac.uk/~cs0gco/fiveusers.doc